Appl. Statist. (1996)
45, *No.* 4, *pp.* 431–446

Estimating Genomic Category Probabilities from Fluorescent *in Situ* Hybridization Counts with Misclassification

By PETER F. THALL†, DEREK JACOBY and STUART O. ZIMMERMAN

University of Texas, Houston, USA

[Received June 1994. Final revision November 1995]

SUMMARY

Fluorescent *in situ* hybridization (FISH) is used in many medical settings to identify the genetic or chromosomal abnormality characterizing a disease. FISH techniques may be used to classify a sample of a patient's cells into genomic categories, one or more of which is associated with the disease. The clinical goal is to determine whether there is a positive proportion of diseased cells in the patient, or to estimate this proportion. Unfortunately, such data are often subject to classification error inherent in FISH methodology. However, when additional data are available from cells of known type, typically from normal subjects, this information may be combined with the patient's data to perform the desired inference while correcting for misclassification. We provide a method for estimating the proportions of cells of each category and testing whether a particular proportion is positive in each of several patients when such background data are available. Our approach is to model the misclassification probabilities, jointly to estimate the model parameters and each patient's cell type proportions by using maximum likelihood and to use this to obtain likelihood ratio tests and confidence intervals. The method is applied to blood cell count data from chronic myelogenous leukaemia patients, where FISH is used to identify the chromosomal translocation characterizing the disease.

Keywords: Classification error; Fluorescent *in situ* hybridization; Leukaemia; Maximum likelihood; Product multinomial

1. Introduction

Fluorescent *in situ* hybridization (FISH) has become a powerful tool for identifying specific regions of the human genome. FISH uses coloured fluorescent markers to determine whether one or more specific deoxyribonucleic acid sequences are present in a chromosomal region (domain) in a cell. An important application of this technology arises in the diagnosis of haematologic diseases which may be characterized by a specific genetic alteration or chromosomal abnormality. FISH probes characterizing the abnormality may be used to classify each of a sample of blood or bone marrow cells taken from a patient into specific categories, one or more of which is associated with the disease. In practice, inference from the resulting multinomial data may be complicated by classification errors that are often inherent in the use of FISH probes. Consequently, one is faced with the problem of testing whether any diseased

†*Address for correspondence*: Department of Biomathematics, Box 237, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA.
E-mail: rex@odin.mdacc.tmc.edu

 0035–9254/96/45431

cells are present in the patient, or of estimating the proportion of diseased cells, on the basis of a multinomial sample with possible misclassification. When additional multinomial data are available from cells of known type, typically cells from normal subjects, this information may be used in conjunction with the patient's data to perform the desired inference while correcting for misclassification. This paper is motivated by such an application.

Seong *et al.* (1994) recently applied FISH to assist in the diagnosis of chronic myelogenous leukaemia (CML) by identifying the translocation between chromosomes 9 and 22 which characterizes CML. The simultaneous use of two FISH probes of different colours for the break points on these two chromosomes associated with this translocation produces two observable genomic domains in normal cells, and three domains in CML cells. As there are non-zero probabilities of missing domains or of seeing domains not actually present, cells with either one or four domains are also observed. Table 1 presents data of this form, taken from Table 1 of Seong *et al.* (1994). We include five CML patients for illustration, although the methods described here accommodate an arbitrary number of patients. Whereas the statistical development is presented in terms of probabilities, the tables give the data and estimates as percentages, as is customary among cytogeneticists.

The general statistical problem is to test whether the proportion of three-domain cells in an individual patient is non-zero and, if so, to estimate this proportion. These inferences may serve in turn as the basis for deciding whether a newly examined patient has CML, or whether a CML patient previously brought into remission has relapsed and requires therapeutic intervention. The statistical problem of determining adequate sample sizes for FISH studies is also very important, since classifying and counting cells is extremely labour intensive.

Various aspects of the problems of estimating multinomial probabilities in the presence of classification error and of evaluating FISH probes have been addressed by several researchers. In a study of sex-mismatched bone marrow transplantation (BMT), Durnam *et al.* (1989) used bone marrow samples from normal males and normal females to test the sensitivity of a Y-chromosome-specific FISH assay and also assessed the ability of the proportion of host cells in the patient measured with this assay to predict relapse or acute graft *versus* host disease. Jenkins *et al.* (1992) discussed the use of interphase and metaphase FISH methods to estimate the proportion of cells with trisomy 8, an abnormality of the eighth chromosome

TABLE 1
Domain counts for normal subjects and five CML patients†

No. of domains (k)	*Normal subjects*	*Counts for the following patients:*				
		1	*2*	*3*	*4*	*5*
1	46 (4.23)	31 (5.20)	22 (5.47)	28 (7.69)	4 (2.04)	38 (8.17)
2	1008 (92.6)	110 (18.5)	376 (93.5)	226 (62.1)	13 (6.63)	402 (86.5)
3	21 (1.93)	453 (76.0)	3 (0.746)	109 (29.9)	174 (88.8)	20 (4.30)
4	13 (1.19)	2 (0.336)	1 (0.249)	1 (0.275)	5 (2.55)	5 (1.08)
n	1088	596	402	364	196	465

†Percentages, out of the total number n of cells classified, are given in parentheses.

associated with certain haematologic disorders, accounting for laboratory error, disagreement between FISH and conventional cytogenetic methods, and interobserver variability. Kibbelaar *et al.* (1993) assessed the ability of three statistical tests to discriminate between test and control FISH probes, while also considering interobserver variability. Sopor and Troilo (1992) developed a test to distinguish true FISH binding sites from non-specific binding sites, by first partitioning chromosomes into bins of equal length and counting the number of binding sites in each bin. Zelen and Haitovsky (1991) studied the asymptotic relative efficiency of tests for comparing two binomial populations when the binary variables are subject to misclassification. Lakshmi and Smith (1993) used power and sample size criteria based on the two-sample binomial test with an inverse sine transformation to evaluate false positive and false negative rates arising in classifying normal and abnormal cells by using flow cytometry.

Our goal here is to use the background data to account formally for misclassification error and thus to obtain more reliable estimates and tests of the cell type proportions in the patient. We do this by first modelling the misclassification probabilities, and then estimating the parameters of the model and the patient's cell type proportions jointly by using maximum likelihood. We provide a likelihood ratio test to determine whether the proportion of three-domain cells in the patient is non-zero and a confidence interval for estimating this proportion. The properties of these methods are examined in the context of the CML application by a simulation study.

2. General Model

For simplicity, we first describe estimation based on background data and data from one patient, and subsequently treat the case of several patients. Let K denote the number of observed cell types. Each patient's data consist of the counts $\mathbf{W} = (W_1, \ldots, W_K)$, where W_k is the number of cells out of n classified that are observed to be of type k, for $k = 1, \ldots, K$. The background data consist of a vector $\mathbf{Z} = (Z_1, \ldots, Z_K)$, from m cells of known type, typically from normal subjects. Our objective is to form inferences about the vector $\pi = (\pi_1, \ldots, \pi_{K-1})$ of the patient's true cell type proportions based on $\mathbf{Z}$ and $\mathbf{W}$. A central point is that the event that a cell is truly of type k is not the same as the event that it is *observed* to be of type k. Denote the probability that a cell is observed to be of type k by ψ_k, for $k = 1, \ldots, K$, with $\psi = (\psi_1, \ldots, \psi_{K-1})$ and $\lambda_{kj} = \Pr(\text{a cell is observed to be of type } k \mid \text{the cell is truly of type } j)$. In general, $\pi \neq \psi$, but

$$\psi_k = \sum_{j=1}^{K} \lambda_{kj} \pi_j, \qquad k = 1, \ldots, K, \tag{1}$$

where $\pi_K = 1 - \pi_1 \ldots - \pi_{K-1}$; ψ_K and λ_{Kj} are defined similarly.

Our estimation scheme requires an explicit representation for each element of the $K \times K$ matrix $\Lambda = (\lambda_{kj})$ of conditional probabilities in terms of a vector θ of parameters, based on a fundamental model for misclassification. The basic idea is to model Λ parsimoniously and to estimate the misclassification parameters θ jointly with π. We apply two such misclassification models in Section 5 to analyse the CML data given in Table 1. In general, we assume that the misclassification

probabilities are the same for all cells, from both normal subjects and CML patients, and that classifications are independent from cell to cell. In particular, this implies that the probability a normal cell is classified to be of type k equals λ_{k2}, and these proportions are the same for all normal subjects. The background data consist of the domain counts from cells known to be normal, i.e. known to have two domains. The counts from normal subject 1 of Seong *et al.* (1994) were excluded from the background data considered here, because a different laboratory worker obtained the data on this initial subject (personal communication from David Seong). Consequently, the misclassification rates for this subject differed from those for the other normal subjects and for subsequent CML patients. This is shown by an exact permutation test (see Mehta and Patel (1983)) for homogeneity among all nine normal subjects' count vectors in Table 1 of Seong *et al.* (1994), which has p-value 0.036, whereas the same test for subjects 2–9 has p-value 0.343.

Joint maximum likelihood estimators (MLEs) of π and θ are obtained from the likelihood $\mathcal{L}(\theta \mid \mathbf{Z}) \times \mathcal{L}(\pi, \theta \mid \mathbf{W})$, where

$$\mathcal{L}(\theta \mid \mathbf{Z}) = \prod_{k=1}^{K} \lambda_{k2}(\theta)^{Z_k} \tag{2}$$

is the multinomial product based on the vector $\mathbf{Z}$ of genomic category counts from m known normal cells and

$$\mathcal{L}(\pi, \theta \mid \mathbf{W}) = \prod_{k=1}^{K} \psi_k(\pi, \theta)^{W_k} = \prod_{k=1}^{K} \left\{ \sum_{j=1}^{K} \lambda_{kj}(\theta)\pi_j \right\}^{W_k} \tag{3}$$

is the likelihood based on the patient's counts $\mathbf{W}$. This is a version of the likelihood considered by several researchers (Viana (1994), section 2). Our emphasis is on parameterization of Λ, and we consider settings in which certain elements of π are 0. Although each cell from each patient is not of known type, the counts $\mathbf{W}$ provide information for estimating the misclassification parameters θ as well as π because the category observation probabilities ψ are functions of Λ and hence θ. In contrast, the background counts $\mathbf{Z}$ do not provide any information about the true category proportions π of the patient; rather they only contain information about the misclassification parameters θ. Thus, if no background data are available, $\mathcal{L}(\pi, \theta \mid \mathbf{W})$ *per se* is not identifiable in both θ and π, i.e. they cannot be separately estimated.

There are two important points here. The first is that π characterizes the individual patient, and therapeutic decisions for that patient will be based on the numerical value of a particular entry of π, such as the proportion π_3 of cancer cells in a CML patient. Thus the primary goal is to obtain an accurate estimate of π. The second point is that θ characterizes a particular FISH procedure performed by an individual cytogeneticist, so both the background data and the patient data must be obtained by the same laboratory worker. A realistic extension is that θ characterizes the classification distribution for all workers in the same laboratory, provided that the particular FISH procedure has been well standardized in that laboratory, as established by a preliminary analysis of interobserver agreement for known cell types. If this is not the case, then background data obtained by one worker in combination with patient data obtained by another worker may produce misleading results. To apply the

methodology proposed here, both the background and the patient data must be obtained from the same worker or standardized laboratory.

In practice, data from several patients will be available. Under our model each patient's data vector contains information for estimating the misclassification parameters. Hence a more reliable estimator of θ is obtained by extending the likelihood (3) to accommodate data from several patients. Denote the data and probability vectors of N patients by $\{(\mathbf{W}_i, \pi_i),\ i = 1, \ldots, N\}$, where $\pi_i = (\pi_{i1}, \ldots, \pi_{i,K-1})$ and $\mathbf{W}_i = (W_{i1}, \ldots, W_{iK})$, recalling that each patient has a different π_i but the misclassification parameters θ are intrinsic to the cell classification process and hence are common to all patients. Denote the probability that cell type k is observed in patient i by

$$\psi_{ik} = \sum_{j=1}^{K} \lambda_{kj}(\theta)\, \pi_{ij}.$$

Then the full likelihood in $(\theta, \pi_1, \ldots, \pi_N)$ is

$$\mathcal{L}(\theta \mid \mathbf{Z}) \times \prod_{i=1}^{N} \mathcal{L}(\pi_i, \theta \mid \mathbf{W}_i) = \mathcal{L}(\theta \mid \mathbf{Z}) \times \prod_{i=1}^{N} \prod_{k=1}^{K} \psi_{ik}^{W_{ik}}. \tag{4}$$

General expressions for the log-likelihood, scores and information matrix are given in Appendix A. To maximize the full likelihood (4) we used the following two-stage iterative procedure:

(a) fix θ and separately maximize each $\mathcal{L}(\pi_i, \theta \mid \mathbf{W}_i)$ in π_i;
(b) fix $(\pi_1, \ldots, \pi_N)$ and maximize the full likelihood in θ.

Good starting values are $\pi_{ij}^0 = W_{ij}/n_i$ with the entries of θ^0 similarly obtained by equating them to the empirical rates in the background data. We do not recommend simultaneous maximization of the full likelihood in all parameters, since we found this method to be numerically unstable, and the problem grows in severity with the number of patients.

3. One Binary Outcome with Misclassification

The simplest application of the general approach is to the case of a single binary outcome subject to misclassification. One example arises from the use of a single probe for the Y-chromosome in sex-mismatched allogeneic BMT, where the donor and patient are of opposite sex. When BMT is used as a therapeutic strategy for haematologic malignancies, such as leukaemia, lymphoma or myelodysplastic syndrome, high dose chemotherapy and possibly radiotherapy are first employed to ablate (eradicate) the patient's bone marrow. In an *allogeneic* transplant, a donor's marrow cells are then introduced into the patient. Any of the patient's original (*host*) bone marrow cells that remain after the transplant may interfere with the process of engraftment, whereby the donor cells repopulate the patient's marrow, and host cells may also contribute to relapse. When the patient and donor are of opposite sex, one way to determine whether host cells remain in the patient and, if so, to estimate the proportions of host and donor cells is to classify a sample of cells from the patient by

their sex chromosomes. Since human male blood cells have one X- and one Y-chromosome whereas female cells have two X-chromosomes, a FISH probe for a genetic domain found only on the Y-chromosome may be used to characterize each cell.

For this application, the relevant probabilities are $\psi = \Pr(\text{observe Y})$, $\pi = \Pr(\text{Y truly present})$ and $\lambda = \Pr(\text{observe Y} \mid \text{Y truly present})$. Since it is impossible to observe the Y-probe if no Y-chromosome is present in the cell, the probability of observing Y is simply $\psi = \lambda\pi$. Here the binomial variable W is the number of the patient's cells observed to have a Y-domain out of n classified, Z is the number of known male cells with an observed Y-domain out of m classified and the likelihood is

$$\mathcal{L}(\pi, \lambda) = (\lambda\pi)^W (1 - \lambda\pi)^{n-W} \lambda^Z (1 - \lambda)^{m-Z}.$$

The MLEs are $(\hat{\pi}, \hat{\lambda}) = (Wm/Zn, Z/m)$ with asymptotic variance–covariance matrix

$$\begin{pmatrix} \dfrac{\pi(1-\lambda\pi)}{\lambda n} + \dfrac{\pi^2(1-\lambda)}{\lambda m} & -\dfrac{\pi(1-\lambda)}{m} \\ -\dfrac{\pi(1-\lambda)}{m} & \dfrac{\lambda(1-\lambda)}{m} \end{pmatrix}.$$

If misclassification is ignored and the incorrect estimator W/n is used for π, this is too small by the factor $\hat{\lambda} = Z/m$. The variance of W/n is not $\pi(1-\pi)/n$ but $\lambda\pi(1-\lambda\pi)/n$, which is smaller than the variance of the MLE $\hat{\pi}$. For example, if $\pi = 0.80$ and $\lambda = 0.95$, corresponding to a 5% misclassification rate, then W/n is an unbiased estimator of $\lambda\pi = 0.76$ rather than 0.80 and has variance $0.182/n$. The estimator of π which accounts for misclassification is asymptotically unbiased, however, and has variance $0.202/n + 0.034/m$.

One *caveat* here, pointed out by a referee, is that, although W/n has mean $\lambda\pi$ and Z/m has mean λ, there is a non-zero probability that $W/n > Z/m$. The numerical solution $\hat{\pi}$ is then greater than 1, which is of course inadmissible. We must be aware of this possibility when n is relatively small or π is close to 1. In such a circumstance, a reasonable solution is to compute a confidence interval for π, using the general method described below based on $\hat{\pi} = 1$ and the above variance estimate for $\hat{\pi}$.

4. Inference

The primary focus in a given application is the proportion π_j of a particular type of abnormal cell. Hence our goals are to test the hypotheses $\pi_j = 0$ *versus* $\pi_j > 0$, or simply to estimate π_j. A likelihood ratio statistic $2(\ln \mathcal{L}_{\text{FULL}} - \ln \mathcal{L}_{\pi_j=0})$ for this test for each patient may be obtained from the likelihood maximized under the full model and maximized assuming $\pi_j = 0$ for that patient. This statistic is distributed approximately as a 50:50 mixture of a point mass at 0 and a χ^2_1-distribution as $n \to \infty$ and $m \to \infty$ (Self and Liang, 1987) because the null value of (π, θ) is on the boundary of the parameter space. In the CML application, there are four categories and π_3 is the patient's proportion of cancer cells. The clinical goal is to determine whether a new patient has CML, or whether a formerly treated CML patient is in relapse. Under each of several assumed models, depending on which actual domain probabilities

may be positive and the underlying misclassification model, we perform likelihood ratio tests of $\pi_3 = 0$ *versus* $\pi_3 > 0$ for each of the patients.

To estimate π_3 we adapt a method first proposed by Ghosh (1979) in the context of constructing an approximate confidence interval for a simple binomial proportion π. Ghosh showed via simulation that, denoting $A = z_\alpha^2/n$ where z_α is the upper $100(1-\alpha/2)$ standard normal percentile and $\hat{\sigma}_n^2 = \hat{\pi}(1-\hat{\pi})/n$, the approximate $100(1-\alpha)\%$ confidence interval

$$\frac{\hat{\pi} + A/2 \mp z_\alpha(\hat{\sigma}_n^2 + A/4n)^{1/2}}{1+A} \tag{5}$$

is superior to the usual large sample interval $\hat{\pi} \mp z_\alpha \hat{\sigma}_n$. This superiority is especially pronounced for very small π, which is important in the present context. Ghosh motivated the improved interval (5) by pointing out that it is based on the shrinkage estimators

$$\tilde{\pi} = \hat{\pi}/(1+A) + 0.5\,A/(1+A)$$

and

$$\tilde{\sigma}_n^2(\tilde{\pi}) = \frac{\hat{\sigma}_n^2 + A/4n}{(1+A)^2}.$$

We adapt this approach to the present problem by using the maximum likelihood estimators of π_3 and $\sigma_n^2(\pi_3)$, obtained under the models accounting for misclassification, in place of $\hat{\pi}$ and $\hat{\sigma}_n^2$ in expression (5). A small simulation study of both the likelihood ratio test and the adapted Ghosh confidence interval, described below in the context of our application, shows that for small values of π_3 the adapted Ghosh interval is substantially more accurate than the usual large sample interval.

5. Analysis of Chronic Myelogenous Leukaemia Data

For the data in Table 1, $k = 1, 2, 3$ or 4 FISH domains may be observed in each cell, with normal and leukaemic cells indexed by $k = 2$ and $k = 3$ respectively. The background data consist of the single vector $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ of observed counts from a sample of cells known to be normal, $m = \Sigma_{k=1}^4 Z_k$ and $\mathcal{L}(\boldsymbol{\theta}|\mathbf{Z}) = \Pi_{k=1}^4 \lambda_{k2}(\theta)^{Z_k}$. The model for π depends on which of three different biological viewpoints is adopted. The first is that only diploid or leukaemic cells can occur, so that one or four domains are observed only as a consequence of technical errors in the FISH process, and $\pi_2 + \pi_3 = 1$. Alternatively, we may assume that all observed categories are possible, i.e. all $\pi_j > 0$. A third possibility, suggested by the data, is that $\pi_4 = 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$. Since there is some disagreement among cytogeneticists about what is possible in these and similar settings, we shall consider all three sets of assumptions.

We consider two models for $\boldsymbol{\Lambda}$, each based on the idea that the misclassification probabilities $\{\lambda_{kj},\ k \neq j \text{ and } j \neq 2\}$ may be obtained from the probabilities $\{\lambda_{k2},\ k \neq 2\}$ for normal cells under the assumption that the conditional probability of wrongly adding or deleting a given number of domains is the same for patient and normal cells. Model 1 for $\boldsymbol{\Lambda}$ is based on the assumptions that any given actual domain is missed with some probability α, an extra domain not actually present is

observed with probability β, the events of missing or adding two or more domains are independent and $\lambda_{kj} = 0$ if $|k-j| > d$, for some integer d. Thus $\theta = (\alpha, \beta)$, $\lambda_{kj} = \alpha^{j-k}$ if $k < j$ and β^{k-j} if $k > j$ and the diagonal elements of Λ are determined by the fact that each column must sum to 1. Under model 1,

$$\Lambda = \begin{pmatrix} 1-\beta-\beta^2-\beta^3 & \alpha & \alpha^2 & \alpha^3 \\ \beta & 1-\alpha-\beta-\beta^2 & \alpha & \alpha^2 \\ \beta^2 & \beta & 1-\alpha-\alpha^2-\beta & \alpha \\ \beta^3 & \beta^2 & \beta & 1-\alpha-\alpha^2-\alpha^3 \end{pmatrix}.$$

The parameters must satisfy the constraints $\alpha + \beta + \beta^2 < 1$ and $\alpha + \alpha^2 + \beta < 1$. Reasonable starting values for computation are $\alpha^0 = Z_1/m$ and $\beta^0 = Z_3/m$, and in general $\pi^0 = \mathbf{W}/n$.

Our second model for Λ is based on the assumptions that at most two extra domains may be observed and at most one may be lost. Specifically, we let $\lambda_{12} = \alpha$, $\lambda_{32} = \beta$, $\lambda_{42} = \gamma$ and $\lambda_{22} = 1-\alpha-\beta-\gamma$, and we assume that $\lambda_{kj} = 0$ if $i < j-1$ or $i > j+2$. This produces model 2, given by

$$\Lambda = \begin{pmatrix} 1-\beta-\gamma & \alpha & 0 & 0 \\ \beta & 1-\alpha-\beta-\gamma & \alpha & 0 \\ \gamma & \beta & 1-\alpha-\beta & \alpha \\ 0 & \gamma & \beta & 1-\alpha \end{pmatrix}.$$

A more symmetric version of this model would have a fourth parameter, say τ, for the probability of missing two domains, with $\lambda_{13} = \lambda_{24} = \tau$. As there is no zero-domain category, and hence no Z_{02} or λ_{02}, such a model is inappropriate here. Starting values for computing the MLEs are $\alpha^0 = Z_1/m$, $\beta^0 = Z_3/m$ and $\gamma^0 = Z_4/m$.

Assume first that each patient cell can have either two or three domains. Thus $\pi_2 + \pi_3 = 1$ and non-zero values of Z_1, Z_3 or Z_4 are due to classification error. The individual patient likelihood (3) is

$$\mathcal{L}(\pi, \theta | \mathbf{W}) = \prod_{k=1}^{4} \{\lambda_{k2}(\theta)(1-\pi_3) + \lambda_{k3}(\theta)\pi_3\}^{W_k}.$$

Table 2 summarizes likelihood ratio tests of $\pi_3 = 0$ *versus* $\pi_3 > 0$ together with point estimates and 99% confidence intervals I_{99} for π_3 for each patient under each misclassification model. The uncorrected estimates and intervals assuming no misclassification, i.e. when $\pi_3 = \psi_3$, are also included for comparison. If a misclassification model is not assumed, however, then $\psi_j = \pi_j$ for each $j = 1, 2, 3$ and 4. Under this approach, the only multinomial model without positive misclassification probabilities which conforms to the observed data is that in which all four π_js are positive. If we do not allow the possibility of misclassification and moreover assume that $\pi_2 + \pi_3 = 1$, then the observation of cells with one or four domains is considered impossible even though such cells are observed. Thus we only consider likelihood ratio tests under the misclassification models, since they allow positive values of all ψ_j.

TABLE 2
Estimates of parameters for two misclassification models and proportion π_3 of leukaemic cells in each of five CML patients; estimates of π_3 assuming no misclassification†

	Estimates for the following patients:				
	1	*2*	*3*	*4*	*5*
Model 1: $\hat{\alpha}$ = 7.05 (0.653), $\hat{\beta}$ = 2.68 (0.461)					
$\hat{\pi}_3$	81.9 (2.09)	0.00 (1.06)	30.6 (2.72)	98.6 (2.46)	2.98 (1.29)
I_{99}	76.2–86.9	0.00–3.62	24.0–37.9	90.7–100	0.309–6.98
p-value‡	0.000	1.000	0.000	0.000	0.002
Model 2: $\hat{\alpha}$ = 6.77 (0.353), $\hat{\beta}$ = 1.37 (0.237), $\hat{\gamma}$ = 0.858 (0.641)					
$\hat{\pi}_3$	80.3 (2.05)	0.00 (0.750)	31.1 (2.65)	95.1 (2.53)	3.25 (1.10)
I_{99}	74.7–85.2	0.00–2.88	24.7–38.2	87.1–100	1.02–6.79
p-value‡	0.000	1.000	0.000	0.000	<0.001
Uncorrected§					
$\hat{\pi}_3$	76.0 (1.75)	0.746 (0.429)	29.9 (2.40)	88.8 (2.25)	4.30 (0.941)
I_{99}	71.2–80.2	0.167–2.70	24.2–36.5	81.7–93.4	2.40–7.34

†Assumes $\pi_2 + \pi_3 = 1$. Standard errors are given in parentheses. I_{99} is an adapted Ghosh 99% confidence interval for π_3.
‡Likelihood ratio test of $\pi_3 = 0$ *versus* $\pi_3 > 0$.
§Assumes no misclassification.

It appears that model 2 gives a better fit to these data than model 1 does. Although models 2 and 1 are not nested, their respective maximized log-likelihoods are -1580.26 and -1606.49. In terms of the observed background normal cell rates of 1.93% for three and 1.19% for four domains (Table 1), the estimates $\hat{\beta} = 1.37\%$ and $\hat{\gamma} = 0.858\%$ under model 2 are closer than the estimates $\hat{\beta} = 2.68\%$ and $\hat{\beta}^2 = 0.072\%$ under model 1.

For patients 1 and 4, both misclassification models give estimates of π_3 that are substantially larger than the uncorrected values. The reverse is true for patients 2 and 5, whose observed proportions of cells with three domains are small. The standard errors of the corrected estimates are larger than those of the uncorrected estimates in all cases. Under either model 1 or model 2, the likelihood ratio tests of $\pi_3 = 0$ *versus* $\pi_3 > 0$ strongly support the null hypothesis for patient 2 and the alternative for each of the other patients. However, the confidence intervals provide more information.

For patient 2, the misclassification models lead to a conclusion that is different from that of the uncorrected model, since the lower confidence bound (LCB) for patient 2 is 0 under both misclassification models, whereas the LCB based on the uncorrected model is positive. Thus, on the basis of a 0.01-level test corresponding to the Ghosh confidence interval, we conclude that $\pi_3 = 0$ for this patient, whereas ignoring misclassification leads to the conclusion that $\pi_3 > 0$, i.e. that patient 2 has a positive proportion of leukaemia cells. A sensitivity analysis in which counts are shifted from the two-domain to the three-domain category for this patient, i.e. the data $\mathbf{W} = (22, 376 - x, 3 + x, 1)$ are analysed for each $x = 3, 4, \ldots$, shows that the 99% LCB is first positive for $\mathbf{W} = (22, 356, 23, 1)$ under model 1 (I_{99} of 0.079%–7.40%), and for $\mathbf{W} = (22, 362, 17, 1)$ under model 2 (I_{99} of 0.166%–6.35%). Thus the models accounting for misclassification require a larger number of three-domain cells to declare a patient to have any leukaemic cells.

To examine the behaviour of the likelihood ratio test of $\pi_3 = 0$ *versus* $\pi_3 > 0$, and to compare the Ghosh and conventional confidence intervals, we performed a simulation study. This was carried out under misclassification model 2 and assuming that $\pi_2 + \pi_3 = 1$. Under the full model with $\pi_3 > 0$, $\psi_1 = \alpha(1-\pi_3)$, $\psi_2 = (1-\beta-\alpha-\gamma)(1-\pi_3)+\alpha\pi_3$, $\psi_3 = \beta(1-\pi_3)+(1-\alpha-\beta)\pi_3$ and $\psi_4 = \gamma(1-\pi_3)+\beta\pi_3$. Thus, under the null model with $\pi_3 = 0$, the MLEs of the misclassification probabilities are the simple multinomial proportions $\hat{\alpha}^0 = (Z_1 + W_1)/(m_1 + n_1)$, $\hat{\beta}^0 = (Z_3 + W_3)/(m_3 + n_3)$ and $\hat{\gamma}^0 = (Z_4 + W_4)/(m_4 + n_4)$. To reflect the observed classification rates in the simulations, we fixed the misclassification parameters at these null empirical values; moreover, the background data vector $\mathbf{Z} = (46, 1008, 21, 13)$ was used throughout. For each case, we generated $W \sim$ multinomial(n, ψ), with $n = 50$, 100 or 200, and π_3 varied along the domain from 0 to 0.20. Each case was simulated 400 times. Fig. 1 presents the resulting power curves for the 0.025-level test based on the χ^2 mixture approximation, i.e. the test which rejects the hypothesis $\pi_3 = 0$ if the likelihood ratio statistic is greater than 3.8415. We also recorded the empirical coverage probabilities and widths of the Ghosh and conventional 95% confidence intervals for these cases and also for $n = 500$ at $\pi_3 = 0.001$, 0.01, 0.05, 0.10, 0.20 and 0.50 to obtain a more complete empirical evaluation of the relative merits of these two procedures. The power functions are graphed in Fig. 1, and the confidence interval coverage rates and widths are given in Table 3.

Since the misclassification models are most useful for true values of π_3 near 0, the power curves indicate that, in the present setting, at least 200 cells must be classified to have a reasonable probability of detecting positive values of π_3 below 0.05 on the basis of the 0.025-level likelihood ratio test. The empirical sizes were 0.028, 0.015 and 0.018 for $n = 50$, 100 and 200 respectively, indicating that the test is somewhat conservative for $n \geqslant 100$. For estimation, the adapted Ghosh confidence interval is

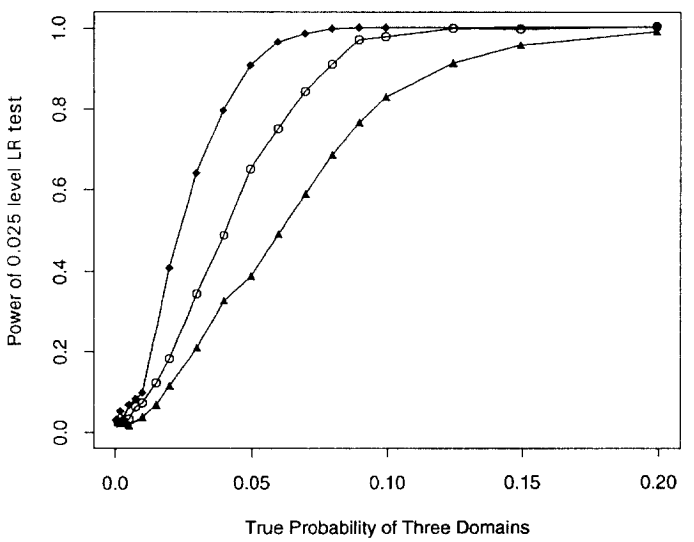

Fig. 1. Power functions of a likelihood ratio test of $\pi_3 = 0$ *versus* $\pi_3 > 0$ under misclassification model 2, assuming $\pi_2 + \pi_3 = 1$ (each point is based on 400 simulated 0.025-level tests): ◆, $n = 200$ cells; ○, $n = 100$ cells; ▲, $n = 50$ cells

TABLE 3
Empirical mean coverage probabilities and widths of Ghosh (1979) and conventional 95% confidence intervals for π_3†

Results for the following true values of π_3:					
0.001	*0.01*	*0.05*	*0.10*	*0.20*	*0.50*
n = 50					
0.948 (0.102)	0.962 (0.109)	0.950 (0.147)	0.935 (0.187)	0.968 (0.233)	0.930 (0.283)
1.00 (0.059)	1.00 (0.066)	0.842 (0.118)	0.908 (0.178)	0.940 (0.238)	0.930 (0.295)
n = 100					
0.935 (0.063)	0.982 (0.070)	0.978 (0.097)	0.960 (0.134)	0.960 (0.169)	0.968 (0.206)
1.00 (0.042)	1.00 (0.050)	0.925 (0.097)	0.915 (0.133)	0.940 (0.171)	0.958 (0.295)
n = 200					
0.985 (0.040)	0.975 (0.047)	0.948 (0.077)	0.968 (0.096)	0.950 (0.120)	0.958 (0.148)
0.995 (0.030)	0.988 (0.037)	0.938 (0.075)	0.965 (0.096)	0.948 (0.121)	0.958 (0.149)
n = 500					
0.980 (0.024)	0.980 (0.033)	0.948 (0.050)	0.945 (0.062)	0.948 (0.077)	0.942 (0.095)
0.993 (0.021)	0.998 (0.029)	0.950 (0.050)	0.960 (0.062)	0.952 (0.077)	0.942 (0.095)

†Based on 400 simulations per case under misclassification model 2 with $\pi_2 + \pi_3 = 1$. For each n, the first row corresponds to the Ghosh intervals and the second row to the conventional intervals. The widths of the intervals are given in parentheses.

far superior to the conventional interval in terms of coverage probability for $\pi_3 =$ 0.001–0.01 or $n = 50$–100, although the Ghosh intervals are slightly wider. Even when n is large or π_3 is closer to 0.50, however, the Ghosh interval is still either slightly more accurate or essentially identical with the conventional interval. These results are analogous to those obtained by Ghosh (1979) in the simple binomial context.

Table 4 presents results that are analogous to those of Table 2, but under the alternative assumption that one or four domains are possible and not purely artefacts of classification error; equivalently that all $\pi_j > 0$. In this case, to allow unconstrained maximization of the likelihood, in the numerical computations we transformed π to ϕ given by $\phi_j = \log(\pi_{j+1}/\pi_j), j = 1, 2, 3$. Under model 1, the Ghosh intervals and likelihood ratio tests are very close to the corresponding values obtained earlier assuming $\pi_1 = \pi_4 = 0$. In general, allowing the possibility of cells with four domains has the effect of decreasing the estimated proportion of three-domain cells under either misclassification model. Under model 1 this leads to a different conclusion for patient 5, since here the 99% interval for this patient now has LCB 0, which implies that this patient may be free of leukaemia cells. The likelihood ratio test p-value, although still small, has increased from 0.0024 to 0.0144 for this patient. The LCB is positive for patient 5 under model 2, however, and the test p-value is very small. Moreover, since the upper bound of the 99% interval is 5.74% under model 1, it seems reasonable to conclude that this patient has a positive proportion of leukaemia cells under either misclassification model or set of assumptions regarding π.

As most values of $\hat{\pi}_4$ in Table 4 are 0 or very small, we consider one more set of assumptions, that $\pi_1 + \pi_2 + \pi_3 = 1$. Analysing the data under this condition with

TABLE 4
Parameter estimates and 99% confidence intervals for $\pi_3 = Pr(CML\ cell)$ in five CML patients†

Parameter	*Results for the following patients:*				
	1	*2*	*3*	*4*	*5*
Model 1: $\hat{\alpha}$ = 4.17 (0.606), $\hat{\beta}$ = 2.48 (0.449)					
$\hat{\pi}_1$	4.50 (0.951)	1.40 (1.38)	5.05 (1.54)	1.80 (1.05)	4.42 (1.49)
$\hat{\pi}_2$	16.1 (1.83)	98.4 (1.72)	65.1 (2.87)	2.80 (2.09)	92.5 (1.91)
$\hat{\pi}_3$	79.4 (2.07)	0.00 (0.980)	29.9 (2.65)	95.2 (2.62)	2.11 (1.13)
$\hat{\pi}_4$	0.00 (0.722)	0.189 (0.258)	0.00 (0.514)	0.199 (1.29)	1.01 (0.502)
I_{99}	73.8–84.4	0.00–3.43	23.5–37.0	87.0–100	0.00–5.74
p-value‡	0.000	1.000	0.000	0.000	0.014
Model 2: $\hat{\alpha}$ = 3.30 (0.339), $\hat{\beta}$ = 1.27 (0.284), $\hat{\gamma}$ = 0.888 (0.541)					
$\hat{\pi}_1$	5.19 (0.956)	4.39 (1.24)	7.19 (1.49)	2.05 (1.06)	7.34 (1.38)
$\hat{\pi}_2$	17.9 (1.69)	93.6 (1.70)	63.2 (2.78)	5.77 (1.91)	89.3 (1.89)
$\hat{\pi}_3$	76.9 (1.84)	0.00 (0.584)	29.6 (2.48)	90.6 (2.35)	3.31 (1.00)
$\hat{\pi}_4$	0.00 (0.533)	0.00 (1.06)	0.00 (0.905)	1.58 (1.18)	0.00 (0.967)
I_{99}	71.8–81.2	0.00–2.50	23.6–36.3	83.2–95.4	1.32–6.61
p-value‡	0.000	0.494	0.000	0.000	<0.001

†Assumes $\pi_j > 0$ for $j = 1, 2, 3, 4$. I_{99} denotes 99% Ghosh confidence intervals for π_3.
‡Likelihood ratio test of $\pi_3 = 0$ *versus* $\pi_3 > 0$.

model 2, summarized in Table 5, produces the same substantive conclusions as obtained by assuming $\pi_4 > 0$, in terms of both the tests and the confidence intervals. The assumption that $\pi_4 = 0$ has the effect of decreasing the standard error of $\hat{\pi}_3$ for all but patient 3 and also produces a more precise estimate of γ. Once again, the analyses indicate that patient 2 is leukaemia free and that the other four patients have a positive proportion of leukaemia cells.

When it is assumed that all $\pi_j > 0$, an alternative method for estimating π is first to estimate ψ empirically with $\hat{\psi} = n^{-1}\mathbf{W}$, and then to maximize $\mathcal{L}(\theta|\mathbf{Z})$ alone in θ. As π and ψ are of the same dimension, the equations $\hat{\psi} = \Lambda(\hat{\theta})\pi$ may then be solved for the $K - 1$ unknowns $\pi_1, \ldots, \pi_{K-1}$. An essential difference between this approach and that given earlier is that here $\hat{\theta}$ is based solely on the background data $\mathbf{Z}$ and does not involve the patients' counts. Other researchers have taken analogous

TABLE 5
Parameter estimates and 99% confidence intervals for $\pi_3 = Pr(CML\ cell)$ in five CML patients†

Parameter	*Results for the following patients:*				
	1	*2*	*3*	*4*	*5*
$\hat{\pi}_1$	4.63 (0.912)	1.29 (0.923)	5.07 (1.32)	1.97 (1.03)	4.35 (1.17)
$\hat{\pi}_2$	15.9 (1.61)	98.7 (1.05)	64.5 (2.64)	2.80 (1.51)	92.5 (1.47)
$\hat{\pi}_3$	79.4 (1.75)	0.00 (0.570)	30.4 (2.50)	95.2 (1.78)	3.18 (0.987)
I_{99}	74.6–83.6	0.00–2.47	24.4–37.2	89.0–98.5	1.24–6.44
p-value‡	0.000	1.000	0.000	0.000	<0.001

†Assumes $\pi_j > 0$ for $j = 1, 2, 3$ and $\pi_4 = 0$, under misclassification model 2 ($\hat{\alpha}$ = 4.24 (0.614), $\hat{\beta}$ = 1.38 (0.354) and $\hat{\gamma}$ = 0.873 (0.241)). I_{99} denotes 99% Ghosh confidence intervals for π_3.
‡Likelihood ratio test of $\pi_3 = 0$ *versus* $\pi_3 > 0$.

approaches to similar problems involving misclassified data. These include a method for correcting for misclassification in matched pair studies (Greenland, 1989), a model for estimation of relative risk in case–control studies with misclassification (Duffy *et al.*, 1989) and a method discussed by Selen (1986). Given a probability vector ψ and non-singular transition probability matrix Λ, however, it is well known that the unique solution π^* to $\psi = \Lambda\pi$ may not be a probability vector (Viana, 1994). As shown below, this approach gives negative values of π_4 for our CML data. We examine the method's application, however, to see why it does not work, since it may be a reasonable alternative to maximum likelihood when a given data set and model for Λ do produce a solution which is a probability vector.

Given Λ and ψ, first express $\psi = \Lambda\pi$ as the system of $K-1$ independent equations

$$\psi_i - \lambda_{iK} = \sum_{j=1}^{K-1} (\lambda_{ij} - \lambda_{iK})\pi_j, \qquad i = 1, \ldots, K-1. \tag{6}$$

Temporarily regard the vectors π and ψ in their $(K-1)$-dimensional forms and let $\boldsymbol{\lambda}_j = (\lambda_{1j}, \ldots, \lambda_{K-1,j})^{\mathrm{T}}$ and $\boldsymbol{\Gamma} = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_K, \ldots, \boldsymbol{\lambda}_{K-1} - \boldsymbol{\lambda}_K)$, so that the matrix form of equation (6) is $\psi - \boldsymbol{\lambda}_K = \boldsymbol{\Gamma}\pi$. The empirical probabilities $\hat{\psi} = n^{-1}\mathbf{W}$ are approximately normal with mean ψ and covariance matrix $\mathbf{V}_\psi$, written $\hat{\psi} \dot{\sim} N(\psi, \mathbf{V}_\psi)$, where $V_{\psi,ij} = \psi_i(1-\psi_i)/n$ for $i = j$ and $-\psi_i\psi_j/n$ for $i \neq j$. The MLEs $\hat{\theta} \sim N(\theta, \mathbf{V}_\theta)$ obtained by maximization of $\mathcal{L}(\theta|\mathbf{Z})$ alone yield $\Lambda(\hat{\theta})$ and $\Gamma(\hat{\theta})$, where $\mathbf{V}_\theta = (E[-\partial^2\{\log \mathcal{L}(\theta|\mathbf{Z})\}/\partial\theta\theta^{\mathrm{T}}])^{-1}$, not the inverse of the matrix $\mathcal{I}_{\theta\theta}$ based on the full likelihood given earlier. Regarding $\hat{\pi} = \Gamma(\hat{\theta})^{-1}\{\hat{\psi} - \boldsymbol{\lambda}_K(\hat{\theta})\}$ as a function of $\hat{\theta}$ and $\hat{\psi}$, under suitable regularity conditions (Serfling (1980), pages 122–124) that are easily satisfied in the present context, and using the fact that $\mathbf{Y}$ and $\mathbf{Z}$ are independent, it follows that

$$\hat{\pi} \dot{\sim} N(\pi, \mathbf{D}_\psi\mathbf{V}_\psi\mathbf{D}_\psi^{\mathrm{T}} + \mathbf{D}_\theta\mathbf{V}_\theta\mathbf{D}_\theta^{\mathrm{T}}), \tag{7}$$

where $\mathbf{D}_\psi = \partial\pi/\partial\psi$ and $\mathbf{D}_\theta = \partial\pi/\partial\theta$. Because of the linearity of the derivative operator,

$$\begin{aligned} \mathbf{D}_\psi &= \boldsymbol{\Gamma}^{-1} \\ \mathbf{D}_\theta &= -\boldsymbol{\Gamma}^{-1}\frac{\partial\boldsymbol{\Gamma}}{\partial\theta}\boldsymbol{\Gamma}^{-1}(\psi - \boldsymbol{\lambda}_K) - \boldsymbol{\Gamma}^{-1}\frac{\partial\boldsymbol{\lambda}_K}{\partial\theta}, \end{aligned} \tag{8}$$

where $\partial\boldsymbol{\Gamma}/\partial\theta = (\partial\boldsymbol{\Gamma}/\partial\theta_1 \ldots \partial\boldsymbol{\Gamma}/\partial\theta_J)$ is the $(K-1)\times(K-1)J$ block matrix of derivatives and D_θ has jth column $-\boldsymbol{\Gamma}^{-1}(\partial\boldsymbol{\Gamma}/\partial\theta_j)\boldsymbol{\Gamma}^{-1}(\psi - \boldsymbol{\lambda}_K) - \boldsymbol{\Gamma}^{-1}(\partial\boldsymbol{\lambda}_K/\partial\theta_j)$, for $j = 1, \ldots, J$. Given a model expressing the entries of Λ as functions of θ, the problem is thus reduced to estimating θ from equation (2), computing $\partial\Lambda/\partial\theta$ and evaluating the asymptotic variance–covariance matrix of $\hat{\pi}$ given by expressions (7) and (8).

The solution vector for patient 1 is $\hat{\pi} = (4.33, 15.38, 82.44, -2.15)\%$, which is not a probability vector. A comparison with the solution in Table 4 for this patient shows that performing unconstrained optimization in the ϕ_js and then transforming back to the probability scale yields $\hat{\pi}_j = 0$, if the numerical value of $\hat{\phi}_j$ is sufficiently large. Although the use of this numerical device in the context of maximum likelihood

estimation precludes numerical estimators which are outside the admissible domain, it is more useful inferentially to provide a confidence interval rather than a point estimate.

On fundamental grounds and aside from numerical issues, however, the likelihood approach based on equations (1)–(4) seems preferable to that based on equations (6)–(8). This is because the purely empirical estimate $\hat{\psi}$ used at the start of the latter approach ignores the probabilistic relationship (1) between ψ and π. Instead, it is based on the implicit assumption that $\mathbf{W}$ follows a multinomial distribution with probability vector ψ which is unrelated to the true category probabilities π. Thus the idea underlying equations (6)–(8) is to begin with an incorrect estimator and then to correct it, whereas the likelihood-based approach accounts for the relationship between ψ and π from the start. Although a general comparison of the empirical merits of these two approaches is beyond the scope of the present paper, it may be the subject of future investigation.

6. Discussion

The models discussed in this paper can be extended to accommodate more general situations. For example, if zero domains are observed in the CML application, then each probability vector has five entries and Λ is 5×5. In this case, for example, $\lambda_2 = (\alpha^2,\ \alpha,\ 1-\alpha-\alpha^2-\beta-\beta^2,\ \beta,\ \beta^2)^T$ under model 1, whereas $\lambda_2 = (\tau,\ \alpha,\ 1-\tau-\alpha-\beta-\gamma,\ \beta,\ \gamma)^T$ under model 2.

In certain applications there may be background data on more than one type of cell. For example, in sex-mismatched BMT the use of FISH probes of different colours for each of the X- and Y-chromosomes allows identification of the complement of sex chromosomes in each cell. In addition to the normal male XY- and female XX-complements, however, various other categories such as X, XXX, XXY and XYY may also be observed owing to laboratory error. In this application there may be six or more observed categories; moreover background data may be obtained from both normal males and normal females. To accommodate such settings in general, let $\mathcal{N}$ denote the set of indices corresponding to known cell types in the background data. In this case $\mathcal{N}$ has two elements, say 1 and 2, identifying known male and known female cells in the background data. Thus the single vector $\mathbf{Z}$ in the earlier formulation is now replaced by $\mathbf{Z}_1$ and $\mathbf{Z}_2$. In general, let $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{Kj})$ denote the background counts from m_j cells known to be of type j for each index $j \in \mathcal{N}$. The component of the likelihood given earlier by equation (2) is now generalized to the product multinomial

$$\mathcal{L}(\theta \mid \mathbf{Z}) = \prod_{j \in \mathcal{N}} \prod_{k=1}^{K} \lambda_{kj}(\theta)^{Z_{kj}}. \tag{9}$$

We have applied this more general method to a data set of this form by using a model for Λ that accounts for missing or observing extra X- or Y-chromosomes. The estimated misclassification rates were too low to affect substantively the estimators of π, however. Hence, numerical details of this analysis are not reported here.

Central to the approach applied in this paper is the assumption that misclassification arises only from errors in the experimental procedure, e.g. staining or micros-

copy, in particular that the probabilities of such errors for a given cell do not depend on its clinical origin. Should pathophysiology affect the misclassification probabilities, however, the use of normal subjects to establish the background misclassification rates may be inappropriate. Without some estimate of the change in the misclassification rates due to disease status, the method employed in this paper could not be applied. Even when the above assumptions hold, interobserver differences must be minimized and laboratory procedures standardized for the background and patient data.

Acknowledgements

The authors thank Michael Siciliano and David Seong for presenting the problem, and also the referee, Associate Editor and Editor for their constructive comments and suggestions on earlier versions of the paper.

Appendix A

From equations (2) and (3), the log-likelihood for a single patient is

$$l(\pi, \theta) = \sum_{j \in \mathcal{N}} \sum_{k=1}^{K} Z_{kj} \log \lambda_{kj}(\theta) + \sum_{k=1}^{K} W_k \log \left\{ \sum_{j=1}^{K} \lambda_{kj}(\theta)\, \pi_j \right\}. \tag{10}$$

Let $\phi_{kj,r} = \partial \lambda_{kj} / \partial \theta_r$. Then the scores are

$$U_{\theta_r} = \frac{\partial l}{\partial \theta_r} = \sum_{j \in \mathcal{N}} \sum_{k=1}^{K} Z_{kj} \frac{\phi_{kj,r}}{\lambda_{kj}} + \sum_{k=1}^{K} \frac{W_k}{\psi_k} \sum_{j=1}^{K} \phi_{kj,r}\, \pi_j, \qquad r = 1, \ldots, J, \tag{11}$$

and

$$U_{\pi_j} = \frac{\partial l}{\partial \pi_j} = \sum_{k=1}^{K} \frac{W_k}{\psi_k} (\lambda_{kj} - \lambda_{kK}), \qquad j = 1, \ldots, K-1. \tag{12}$$

The asymptotic variance–covariance matrix of the MLEs $(\hat{\pi}, \hat{\theta})$ is the inverse of the information matrix

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\pi\pi} & \mathcal{I}_{\pi\theta} \\ \mathcal{I}_{\theta\pi} & \mathcal{I}_{\theta\theta} \end{pmatrix},$$

where the submatrices have elements

$$\mathcal{I}_{\pi_j, \pi_k} = n \sum_{i=1}^{K} \psi_i^{-1} (\lambda_{ij} - \lambda_{iK})(\lambda_{ik} - \lambda_{iK}), \tag{13}$$

$$\mathcal{I}_{\theta_r, \theta_s} = \sum_{j \in \mathcal{N}} m_j \sum_{k=1}^{K} \left(\frac{\phi_{kj,r} \phi_{kj,s}}{\lambda_{kj}} - \sigma_{kj,rs} \right) + n \sum_{i=1}^{K} \psi_i^{-1} \left(\sum_{j=1}^{K} \phi_{ij,r} \sum_{k=1}^{K} \phi_{ik,s} - \sum_{j=1}^{K} \sigma_{ij,rs}\, \pi_j \right) \tag{14}$$

and

$$\mathcal{I}_{\pi_j, \theta_r} = n \sum_{i=1}^{K} \left\{ \psi_i^{-1} (\lambda_{ij} - \lambda_{iK}) \sum_{k=1}^{K} \phi_{ik,r} \pi_k - (\phi_{ij,r} - \phi_{iK,r}) \right\}, \tag{15}$$

for $1 \leqslant j \leqslant k \leqslant K-1$, $1 \leqslant r \leqslant s \leqslant J$ and $\sigma_{kj,rs} = \partial^2 \lambda_{kj}/\partial\theta_r \partial\theta_s$.

To accommodate several patients, indexed by $i = 1, \ldots, N$, the scores may be obtained by substituting $\{W_{ik}, \psi_{ik}, \pi_{ik}\}$ for $\{W_k, \psi_k, \pi_k\}$ in equations (11) and (12), summing over i in the second double sum in the expression for U_{θ_r} in equation (11) and noting that U_{π_j} is generalized to $U_{\pi_{ij}}$ for $j = 1, \ldots, K-1$ and $i = 1, \ldots, N$. The information matrix may be obtained from the facts that

$$\frac{\partial^2 l}{\partial\theta_r \partial\theta_s} = \frac{\partial^2 l(\theta|\mathbf{Z})}{\partial\theta_r \partial\theta_s} + \sum_{i=1}^{N} \frac{\partial^2 l(\pi_i, \theta|\mathbf{Y}_i)}{\partial\theta_r \partial\theta_s}$$

and $\partial^2 l/\partial\pi_{ij}\partial\pi_{i'j'} = 0$ for $i \neq i'$, so that $\mathcal{I}_{\pi\pi}$ is the $(K-1)N \times (K-1)N$ block diagonal matrix $\text{diag}(\mathcal{I}_{\pi_1\pi_1}, \ldots, \mathcal{I}_{\pi_N\pi_N})$ and $\mathcal{I}_{\pi\theta} = (\mathcal{I}_{\pi_1\theta} \ldots \mathcal{I}_{\pi_N\theta})$.

References

Duffy, S. W., Rohan, T. E. and Day, N. E. (1989) Misclassification in more than one factor in a case–control study: a combination of Mantel–Haenszel and maximum likelihood approaches. *Statist. Med.*, **8**, 1529–1536.

Durnham, D. M., Anders, K. R., Fisher, L., O'Quigley, J., Bryant, E. M. and Thomas, E. D. (1989) Analysis of the origin of marrow cells in bone marrow transplant recipients using a Y-chromosome-specific in situ hybridization assay. *Blood*, **74**, 2220–2226.

Ghosh, B. K. (1979) A comparison of some approximate confidence intervals for the binomial parameter. *J. Am. Statist. Ass.*, **74**, 894–900.

Greenland, S. (1989) On correcting for misclassification in twin studies and other matched-pair studies. *Statist. Med.*, **8**, 825–829.

Jenkins, R. B., Le Beau, M. M., Kraker, W. J., Borell, T. J., Stalboerger, P. G., Davis, E. M., Penland, L., Fernald, A., Espinosa, R., Schaid, D., Noel, P. and Dewald, G. W. (1992) Fluorescence in situ hybridization: a sensitive method for trisomy 8 detection in bone marrow specimens. *Blood*, **79**, 3307–3315.

Kibbelar, R. E., Kok, F., Dreef, E. J., Kleiverda, C. J., Cornelisse, C. J., Raap, A. K. and Kluin, P. M. (1993) Statistical methods in interphase cytogenetics: an experimental approach. *Cytogenetics*, **14**, 716–724.

Lakshmi, D. V. and Smith, W. K. (1993) Comparing proportions in the presence of false positive and false negative instrument sorting errors. *Biometrics*, **49**, 639–641.

Mehta, C. R. and Patel, N. R. (1983) A network algorithm for performing Fisher's exact test in r × c contingency tables. *J. Am. Statist. Ass.*, **78**, 427–434.

Selen, J. (1986) Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *J. Am. Statist. Ass.*, **81**, 75–81.

Self, S. G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Ass.*, **82**, 605–610.

Seong, D. C., Song, M. Y., Henske, E. P., Zimmerman, S. O., Champlin, R. C., Deisseroth, A. B. and Siciliano, M. J. (1994) Analysis of interphase cells for the Philadelphia translocation using painting probe made by inter-alu PCR from a radiation hybrid. *Blood*, **83**, 2268–2273.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Soper, K. A. and Troilo, P. (1992) Exact tests for *in situ* hybridization experiments. *J. Am. Statist. Ass.*, **87**, 78–83.

Viana, M. A. G. (1994) Bayesian small-sample estimation of misclassified multinomial data. *Biometrics*, **50**, 237–243.

Zelen, M. and Haitovsky, Y. (1991) Testing hypotheses with binary data subject to misclassification errors: analysis and experimental design. *Biometrika*, **78**, 857–865.